DOCUMENT RESUME

ED 147 807 CS 203 735

AUTHOR	Suhor, Charles
TITLE	Mass Testing in Composition: Is It Worth Doing Badly?
INSTITUTION	New Orleans Public Schools, La. Div. of Instruction.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
PUB DATE	Jun 77
NOTE	27p.; A report on "A Building-Centered Testing Program in Writing"
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	*Composition Skills (Literary); *English Instruction; *Evaluation Methods; Middle Schools; Performance Based Education; Rating Scales; Secondary Education; Testing; *Testing Problems; *Testing Programs; Test Reliability
IDENTIFIERS	Diederich (Paul B); Elementary Secondary Education Act Title IV; Louisiana (New Orleans)

ABSTRACT

In attempting to meet school-board mandates for competency-based testing in composition, educators must devise the most acceptable testing programs they can. This paper describes a design (the Paul Diederich system) for testing students' writing skills, which yields statistically reliable data on individual students, and reports on a New Orleans project in which the Diederich system was implemented in one middle school, three junior high schools, and three senior high schools. The paper explains the program procedures, which involve the collection of four writing samples from each student and the rating of the papers by eight teachers during a school year; points to some useful data that emerged from the testing program, and to its effectiveness as an inservice tool; and notes problems in administering it, which led to the conclusion that the total design is not practical for mass testing purposes. THe paper then shows how a modified version of the Diederich system was devised for use in New Orleans schools. In the modified design, the total student population in each school is tested only once. Test scores are used, in conjunction with teacher recommendations, as the basis for placing students with poor writing skills in remedial programs. (GW)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED147807

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

MASS TESTING IN COMPOSITION

IS IT WORTH DOING BADLY?

Report on "A Building-Centered Testing Program in Writing" · New Orleans Public Schools · June, 1977 · Charles Suhor, Supervisor of English, Project Director

CS203735

Participating Schools and English Department Heads

Abramson Senior High School	Kurt Leonhard, Department Head
Brother Martin High School	Brother Francis David, Department Head
Carver Senior High School	Ms. Lucille Simms, Department Head
Nicholls Senior High School	Ms. Tanya Menendez, Department Head
Colton Junior High School	Ms. Willie Topps, Department Head
Kohn Junior High School	Ms. Shirley DeJan, Department Head
Lawless Junior High School	Ms. Vickie Gallo and Ms. Betty Bloch, Department Heads
Livington Middle School	Ms. Luevenia Hughes and Ms. Bonnie Payton Department Heads

The program described in this publication was funded through an ESEA grant, Title IV, Part C. However, the opinions expressed herein do not necessarily reflect the position of the U.S. Office of Education or the administering agencies, and no official endorsement should be inferred. This publication may be reproduced without obtaining permission.

Department of Resource Services
Division of Instruction and Child Advocacy
New Orleans Public Schools

Dr. Gene Geisert, Superintendent

Orleans Parish School Board: Mr. Lloyd Rittiner, President
Dr. Mack J. Spears, Vice-President
Mr. Harwood "Woody" Koppel
Mrs. Rose Loving
Dr. William D. Reeves

Project Director's Report

MASS TESTING IN COMPOSITION: IS IT WORTH DOING BADLY?

A Report On "A Building-Centered Testing Program in Writing," New Orleans Public Schools; Funded by the U.S. Office of Education, ESEA, Title IV, Part C.

Charles Suhor, Supervisor of English

Long before the national fever for accountability and competency testing, English teachers and writing specialists were concerned about the problems of evaluating student writing.

The reliability factor in teachers' ratings of student papers has been a persisting problem in assessing student writing skills. Individual English teachers vary greatly in what they value in student writing. The same essay, corrected by five English teachers, might yield a wide range of grades from A to F. Also, an individual teacher's rating of one student paper might change greatly after a short time span has elapsed between the first and second ratings.

Beginning in the 1960's, the College Entrance Examination Board (CEEB) developed ways of achieving inter-rater reliability in evaluation of student writing. Reliable scoring of the CEEB English Composition Test involves training raters to come to a concensus on what constitutes, roughtly speaking, superior papers, above-average papers, below-average papers, and poor papers. The Writing Tests developed in the late 1960's by National Assessment of Educational Progress (NAEP) also call for student writing samples, and the rating system is similar to the CEEB model. Because NAEP has tested a sweeping national sample of students at age 9, 13, and 17, -- a far broader student population than the college-bound high school students taking the CEEB test -- the NAEP results have been of particular interest to educators and the public at large.

At the very best, though, the CEEB and NAEP tests only scratched the surface of the writing assessment problem. Neither design deals with the fact that on any given day, a student might be writing more poorly than s/he normally writes. An upset stomach, problems at home, and testing jitters are some of the many factors that might cause a student to perform badly on a given day. Reliability of student writing scores depends not only on the number of trained readers used, but also on

the number of different essays written by the same student. Student performance on the CEEB test and the NAEP test (or on any other single test) is not suited for diagnosis and prescription of individual student deficiencies, or for individual placement in remedial or advanced programs, or for deciding whether or not a student should be awarded a diploma. To arrive at such crucial instructional and promotional decisions in a manner that is reliable, responsible, and fair to the individual student, far more information is needed than these tests can provide.

The enormous expense of scoring student writings under the CEEB and NAEP designs is another problem. Teachers must be hired and trained to do the ratings. CEEB's student population consists of self-selected, college-oriented students, while NAEP typically uses sampling techniques to gather information about large populations rather than individual students. This is costly enough, but it is only a fraction of the potential expense involved in testing the writing skills of every student in a school district or a state.

Use of an objective test of student writing skills (e.g., multiple choice items) involves the same shortcomings as taking a single writing sample, with the added absurdity of testing students' writing ability without asking them to write. Morever, the standardized objective test is likely to result in less actual writing instruction, because teachers tend to gear instructional programs towards testing systems. Mindful of this, the members of a recent Louisiana State Department of Education Writing Conference on Minimum Proficiency unanimously recommended that "any test of writing proficiency include a sample of the student's writing."

Unfortunately, the numerous State Legislatures and School Boards that are mandating competency-based testing have not typically concerned themselves with cost implications, or with the inability of testing technology to implement their mandates in ways that are fair to the individual student. A National School Boards Association study described board members as "not sensitive to . . . or accepting of" mounting protests against standardized testing systems. Educational administrators have not been overly zealous in advertising the sad state of the art of testing, either, except within their all-too-insular professional organizations, like the National Education Association (NEA), Association for Supervision and Curriculum Development (ASCD), National Association for Elementary School Principals (NAESP), and National Council of Teachers of English (NCTE) -- all of which have expressed grave concern about the inadequacy of existing testing systems.

The Diederich System -- A Usable Design?

There is one relatively inexpensive design for testing of student writing skills which utilizes multiple writing samples over a period of time and yields statistically reliable data on individual students. The Paul Diederich system calls for four writing samples from each student, rated independently and anonymously by eight teachers during a school year. The design, though somewhat complex, yields good data -- including a rank in class and a rank in school for each student, as well as comparisons of student performances at each grade level.

The Diederich system has an excellent reputation in the profession. Diederich, a pioneer researcher at Education Testing Service, is a humanist and innovator with unimpeachable credentials. His system is highly esteemed by professional researchers, but it is geared to decision-making about program development and student placement in everyday school situations. Diederich visited dozens of schools explaining his design during the 1960's, and he gave a detailed description of it in Measuring Growth in English (NCTE, 1974).

Yet no large-city system has attempted to use Diederich's design on a mass testing basis. It was thought that a test of workability of Diederich's system in a big-city situation like New Orleans Public Schools might point to some interesting future directions for mass testing in writing. In Fall of 1976 an ESEA Title IV program called "A Building-Centered Testing Program in Composition" was begun in New Orleans, focusing on use of the Diederich system in seven District II public schools and one non-public school, Brother Martin High School. District II schools are highly varied, composed of over 10,000 students in three senior high schools, three junior high schools, and a middle school. Abramson High School and Livingston Middle School might be described as mixed suburban, drawing from New Orleans' more affluent eastern area, while the other schools (Colton, Kohn, and Lawless Junior Highs; Carver and Nicholls Senior Highs) draw almost totally from Black working class and welfare populations. Brother Martin is a predominately white, affluent college preparatory Catholic High School of about 1,400 students.

Trying the Diederich system, with its various refinements and complexities, in such a large and varied array of schools seemed like an interesting prospect. The original proposal stated that

the Diederich design is building-centered, curriculum relevant, and inclusive of a wide range of compositional concerns. There is no reason why the Diederich plan cannot be used at the middle, junior, and senior high school levels in a large inner city school system, except that logistical aspects of the design tend to discourage its use when paraprofessional assistance is not available at the building level.

Program Description

The New Orleans program followed the Diederich design rather closely, giving four tests in the participating schools in November, January, March, and May. Student anonymity would be achieved by use of the Student Identification Slip (Figure 1), thus assuring objectivity

Figure 1

Student Identification Slip

Six Digit Number _____

Name _____ Total Score _____

Teacher's Name _____

Grade Level _____ Class (Period) _____ Date _____

in teacher ratings. Each student paper was to be rated twice via Diederich's Rating Scale (Figure 2), an instrument that includes numerous aspects of writing. Teachers would be trained in use of the Reader's Scale, with the goal of spending not more than two minutes on rating an individual paper. This is called "holistic" rating. It involves rapid, impressionistic evaluation of a paper rather than labored, analytical rating. The Educational Testing Service research on rater reliability in the 1960's revealed that multiple ratings based on overall impressions were the best means of achieving inter-rater reliability.

Figure 2

Reader's Scale

Reader No. _____ Paper 6 Digit No. _____

(circle one)	Low		Middle		High	
Ideas	2	4	6	8	10	
Organization	2	4	6	8	10	
Wording	1	2	3	4	5	
Flavor	1	2	3	4	5	
						_____
Usage	1	2	3	4	5	
Punctuation	1	2	3	4	5	
Spelling	1	2	3	4	5	
Handwriting	1	2	3	4	5	
						_____
					sum	_____

Student scores on each test were to be recorded on a Student Writing Record (SWR) (Figure 3), a longitudinal form designed for the New Orleans program. At the end of the session each student's rank in class and rank in school (expressed in quartiles) were to be computed and entered on the SWR. The Quartile Chart (Figure 4) would also be computed at the end of the year to give a picture of the relative quality of student performance at each grade level.

Aside from the double ratings on each student paper, the program's greatest complexity is in the sorting of Student Identification Slips after each round of testing, the recording of student scores on the Student Writing Record, and the computation of final results. This was to be accomplished in five of the schools by hiring paraprofessionals who worked under the direct supervision of English department heads. In two other schools (Carver Senior High and Lawless Junior High) student assistants would do these tasks; at Brother Martin High School, parent volunteers would be used. Since paraprofessionals cost money and funding is only

Figure 3

Holden, Barry M.
last name first initial

STUDENT WRITING RECORD

GRADE 7 19 76 - 77 Session	GRADE 8 19__-__ Session	GRADE 9 19__-__ Session	GRADE 10 19__-__ Session	GRADE 11 19__-__ Session	GRADE 12 19__-__ Session
1st test 38	1st test _____	1st test _____	1st test _____	1st test _____	1st test _____
2nd test 43	2nd test _____	2nd test _____	2nd test _____	2nd test _____	2nd test _____
3rd test 45	3rd test _____	3rd test _____	3rd test _____	3rd test _____	3rd test _____
4th test 39	4th test _____	4th test _____	4th test _____	4th test _____	4th test _____
Final Score 163	Final Score _____	Final Score _____	Final Score _____	Final Score _____	Final Score _____
Rank in class Q4	Rank in class _____	Rank in class _____	Rank in class _____	Rank in class _____	Rank in class _____
Rank in school Q4	Rank in school _____	Rank in school _____	Rank in school _____	Rank in school _____	Rank in school _____

renewable for two years, it seemed important to discover whether use of students and volunteers was feasible, or whether the program was tied to the expense of paraprofessional assistance.

Topics for each of the tests were to be student-centered, expository topics rather than academically-oriented topics or genre assignments (e.g., short stories or poems). If a writing tests depends on the student's use of particular academic information or a specialized skill (like analyzing or writing a poem), then the student who happens not to have the information or who lacks the narrowly defined skill is doomed to poor performance. The idea was to test students' ability to organize, in written form, ideas and feelings known to them, and not to survey student information or specialized skills. English Department Heads would submit topics to the Project Director, usually after discussing possible topics with English department members in their schools.

The program is called "building-centered" because its essential goal is to generate information for use within each school. Although considerable guidance from the Project Director is needed in setting up the program, the school's English department head is essentially the in-house co-ordinator of the program after the teachers have been trained in use of the Rating Scale and the first round of testing has been completed. It was anticipated that decisions about English program development and individual student placement could be made with confidence on the basis of the rich data gathered in the program.

Program Results: What Went Well

The data for each school came in, and it was all very interesting. At informal interpretive sessions, English faculties (and often principals) discussed implications of the data and in some instances made initial plans for program development. The sample Quartile Chart shown in Figure 4 suggests predictable growth from grades 7 to 8, with 50.5% of the seventh graders in the upper half of the school and 57% of the eighth graders in the upper half. Note, though, that only 43.9% of the ninth graders were in the upper half. Program development might be needed at that level, or some unusual qualities might exist in the ninth grade population, or perhaps both.

Some teachers pointed out that many bright students leave the school after grade 8 to attend a highly prestigious alternative high school that stresses a college preparatory program in grades 9-12. Others pointed to their intuition that the current ninth graders were an atypically weak group. Although the latter reason of smacks of rationalization, it was supported by the fact that relatively few ninth graders were on the

Figure 4

SCHOOLWIDE QUARTILE CHART

Number of Students: 288

Quartile Cut-Off Scores	Grade ___	Grade 7	Grade 8	Grade 9
Q 1: 400-228	# %	# 20 20.6 %	# 28 30.1 %	# 25 25.5 %
Q 2: 227-198	# %	# 29 29.9 %	# 25 26.9 %	# 18 18.4 %
% in Top Half		50.5%	57.0%	43.9%
Q 3: 197-166	# %	# 22 22.7 %	# 21 22.6 %	# 29 29.6 %
Q 4: 165-80	# %	# 26 26.8 %	# 19 20.4 %	# 25 26.5 %
Total # and % at each grade level	# 100%	# 97 100%	# 93 100%	# 98 100%

school Honor Roll during this session. The Diederich system cannot reinforce grade-level stereotyping, because teachers do not know the grade levels of students' papers when they do the rating.

The Diederich system's "grade-level blindness" augurs well for the 20.6% of the seventh graders and the 30.1% of the eighth graders who were in the top quartile, in direct competition with the best ninth graders. These students can be encouraged to develop their extraordinary talents more intensively by working on student publications, doing special writing projects, acting as peer editors, etc. Students in the lowest quartile can be given special help in writing through individualized programs and materials.

Since the student's rank in class is given in the rather blunt terms of quartile rankings, further refinement is provided by the Schoolwide Quartile Cut-Off Scores (left hand column of Figure 4, Quartile Chart) and by the grade-level cut-off scores on the Rank-in-Class Chart (Figure 5). For example, a seventh grader like Barry in Figure 3,

Figure 5 RANK-IN-CLASS CHART

FOR GRADE 7

97 Total number of students tested in grade 7

24+ Number of students at grade 7 in each quartile

Quartile Cut-Off Scores, grade 7 (Steps 8 and 9)
Q 1: 400-223
Q 2: 222-202
Q 3: 201-164
Q 4 163-80

with a final score of 163 on the Student Writing Record, might be disappointed that he is in the lowest-performing group (Quartile Four) in both the school and the class. But the teacher can show him that he is only three points away from the third quartile in the school population (cut-off score 166) and only one point from the third quartile in the seventh grade population (cut-off score 164). Conversely, a ninth grader with a score of 200 has barely squeezed into the top half of both the school population and the ninth grade class. (Rank-in-Class Charts are provided for each grade level.)

The students should be told that no generalizations can be made about improvement or lack of improvement between tests one and four, because on any given test day the student might have found the topic less stimulating, might have been feeling ill, etc. The only significant score is the final one, which is based on multiple trials and ratings.

Most schools showed more or less predictable developmental patterns between grade levels on the Schoolwide Quartile Chart, with a few deviations that stimulated interesting interpretive comments. Some charts, though, showed disturbing anomalies, forcing English teachers to do some hard thinking about their programs. Also, many students were surprised by their rankings, which frequently gave recognition to low-profile talent and stripped the halo from overachievers. Such are the benefits of anonymous ratings.

The teachers recognized the value of student anonymity and multiple samples. Seventy-nine percent responded favorably to the evaluation item dealing the "fairness to the student" in terms of "administering, rating, and recording scores in the program."

Another highly successful aspect of the program was the one-day Teacher Training Sessions, during which teachers first learned how to use the Reader's Scale. Actual student papers from the school's first round of testing were used, and teachers enjoyed comparing ratings to see if they came within Diederich's required 10-point range of agreement. There were good discussions about the meanings of the categories on the scale ("Organization," "Flavor," etc.), and teachers were often surprised to learn that their peers held rather exotic views about what is important in the teaching and evaluation of student writing.

Although there was some resistance to the program in general (more about that later), 98% of the teachers reported an understanding of the rating instrument in the post-training evaluation. In the final evaluation, 100% claimed to understand the instrument. Seventy-four percent felt after the training session that they could rate a student's composition in approximately two minutes. This went up to 82% in the

final evaluation. The results here give hope that holistic rating might be used extensively, along with techniques like peer editing, as one way of coping with the age-old problem of finding time to evaluate student writing.

Student-centered writing topics appeared to get a boost from the program, with 72% of the teachers claiming that the topics were appropriate. This feeling was strongest in the junior high and middle schools, where motivation seems to be a strong factor in existing writing programs. The topics were seen as inappropriate by one department head in a highly academic school, where students were trained almost exclusively in literature-related writing in the regular program. (Sample Topics: My Best Friend Is My Best Friend Because. . .; Something I'd Like to See Changed in This School; I Really Get Angry When . . .; If I Had Three Days to Live).

The entire program was a sustained sub rosa in-service exercise. Excellent teachers discovered a quick, interesting rating device. Average teachers gained a new consciousness of composition as they considered hundreds of student papers in light of the categories on the scale. And teachers who had seriously neglected composition instruction were, mirabile dictu, brought into a mainstream of ideas about composition instruction.

A final benefit of the program was the department leadership that emerged in most of the schools. The structure and the status of department chairmanships in New Orleans Public Schools have traditionally been extreemly poor. Department heads are granted no extra preparation time, but they get an annual stipend of $400.00. Under such conditions, the development of leadership roles depends largely on self-motivated department heads and specific programs around which leadership qualities can be developed. The Building-Centered Testing Program required a genuine instructional perspective, skill in human relationships, ability to locate and supervise paraprofessional aids, and endurance in juggling many, many details. The program was a challenge to the experienced department leaders and a baptism of fire for newcomers.

Program Results: What Went Badly

The positive aspects of the program were impressive and gratifying, but the problems of administering it on a large-scale basis were enormous. An individual school with a stable, highly professionalized faculty, good student attendance, adequate departmental organization, and full administrative support would no doubt find the Diederich system superb. But large-scale testing touches on all kinds of situations, and the imponderables in Diederich's system multiply as more and more schools are involved. If a single teacher in a large school is late in returning Reader's Scales to the department head, the entire process of double ratings is frustrated in that school, and the sorting and recording tasks are thrown off rhythm -- with worrisome implications for the revamping of paraprofessionals' schedules. This happened in two schools in the first round of testing.

One of the most talented department heads was promoted to a supervisory position in another school just after the first round, necessitating the rapid-training of a new department leader. Another took an emergency leave because of illness in the family. The project's excellent co-director, Betty Monroe, was transferred from her original job in District II to a central office position. Her participation became fragmentary, focusing mainly on assistance in the Teacher Training Sessions.

One senior high school had a core of teachers who expressed great annoyance over a testing program from the beginning, and raised questions about whether the results would be used as part of the school system's teacher evaluation plan. There were rumblings about referring the matter to the union -- certainly a valid recourse for teachers with doubts about the use of test results -- but the handful of disgruntled teachers apparently did not elect to use institutional channels for grievances, preferring instead to delay the program through non-cooperation at the building level.

These problems were not mere flukes. Administrators and supervisors will recognize such complications as common occurrences in the agitated beehive of a big city administration. A combination of logistical and morale-related snafus, then, prompted the decision to use single rather than double teacher ratings in rounds two through four of the testing. This change in procedure diminished the statistical reliability of the students' final scores. But the operation of the program became manageable, and department heads reported greater administrative facility and more positive teacher attitudes toward the project.

It is interesting that most teachers (79%) did not choose to substitute the four writing tests for regular classroom writing assignments, even though they were invited to do so and most of them (72%) considered the tests topics appropriate. The rating of tests in the program was

therefore a burden over and above regular writing assignments for many teachers, especially those (86%) who reported an existing composition load of two paragraph-length assignments a month. It is not surprising that 92% of the teachers stated that the program involved either "much" extra work, or "a great deal" of extra work in the evaluation.

The teachers' reluctance to substitute the tests for regular assignments might be related to the limited possibilities for feedback. Some schools offered total feedback to students after each round of testing. Others offered partial feedback, while still others -- pointing to the complexity of the sorting and recording system -- provided no formal means of feedback until the end of the program. Only 47% of the teachers used information gained from their students' tests papers in regular classroom instruction -- another flaw in the program that appears to be related to its logistic complexity.

The use of student assistants also proved problematical. Department heads with no additional release time could not supervise students (especially at junior high age) closely enough to assure efficient sorting and recording. Nevertheless, the senior high school using student help continued to do so until the final round, when computation of individual student scores and completion of the Quartile Chart was handled with paraprofessional assistance.

The department heads, who were administering the program at the building level, unanimously described it as either "extremely complicated" or "somewhat complicated". Only 50% of the teachers, who were insulated from the program's more niggling complexities, characterized the program as complicated.

The appeal of the Diederich system is twofold. The Reader's Scale is a fascinating instrument, an excellent vehicle for bringing teachers together to talk about the teaching and evaluation of their students' writing.

The data yielded by the Diederich system is good data, superior to the skimpy information provided on individual students through standardized tests. The New Orleans Public Schools experience indicates that the rating scale is a valuable in-service tool, but that the total design, despite its aesthetic excellence and its research respectability, is not practical for mass testing purposes.

Still, state and local school systems are mandating tests of writing skills. Is a compromise design possible, one which makes use of the most workable aspects of the Diederich system without getting bogged down in its details?

A Modified Design: What Do You Do With A Mandate?

While the Building-Centered Testing Program was in operation in District II in New Orleans Public Schools, the School Board mandated that citywide testing in writing for grades 7-9 be developed, along with remedial courses for students who fail to pass the test. The impracticality of the Diederich design in its original form forced consideration of a compromise design of some sort. A continuation of the ESEA grant was proposed as outlined in the model below (Figure 6).

Any testing problem that deals with placement of students, rewarding or withholding of credit, or the granting of a diploma is also a moral problem. The modified New Orleans design was structured as an attempt to meet the Board's mandate in a way that is administratively feasible, yet as fair as possible to individual student whose educational path is being set, at least temporarily, by less than adequate data. In the modified design the most successful aspects of the Diederich system were merged with relatively simple administrative procedures that include certain safeguards against overdependence on a single test.

The design calls for half-day orientation meetings in September with English department heads in newly participating junior high and middle schools. Details of program structure and responsibilities of department heads and teachers are explained, and department heads are given brief practice sessions with the Reader's Scale. School testing dates are set, as are deadlines for submitting student-centered topics to the Project Director.

In early October, schools administer the first writing test, using a standard set of instructions prepared by the Project Director. Student anonymity is obtained by having students write, on the backs of their papers, the following: name, grade level, English teacher's name, class (period), date. This procedure sets up a grade-level categorization of students and eliminates logistical problems related to the sorting and re-sorting of Student Identification Slips. The design does permit grade-level stereotyping by the raters. But most school schedules are set up by grade levels, and the grade-level categorization facilitates

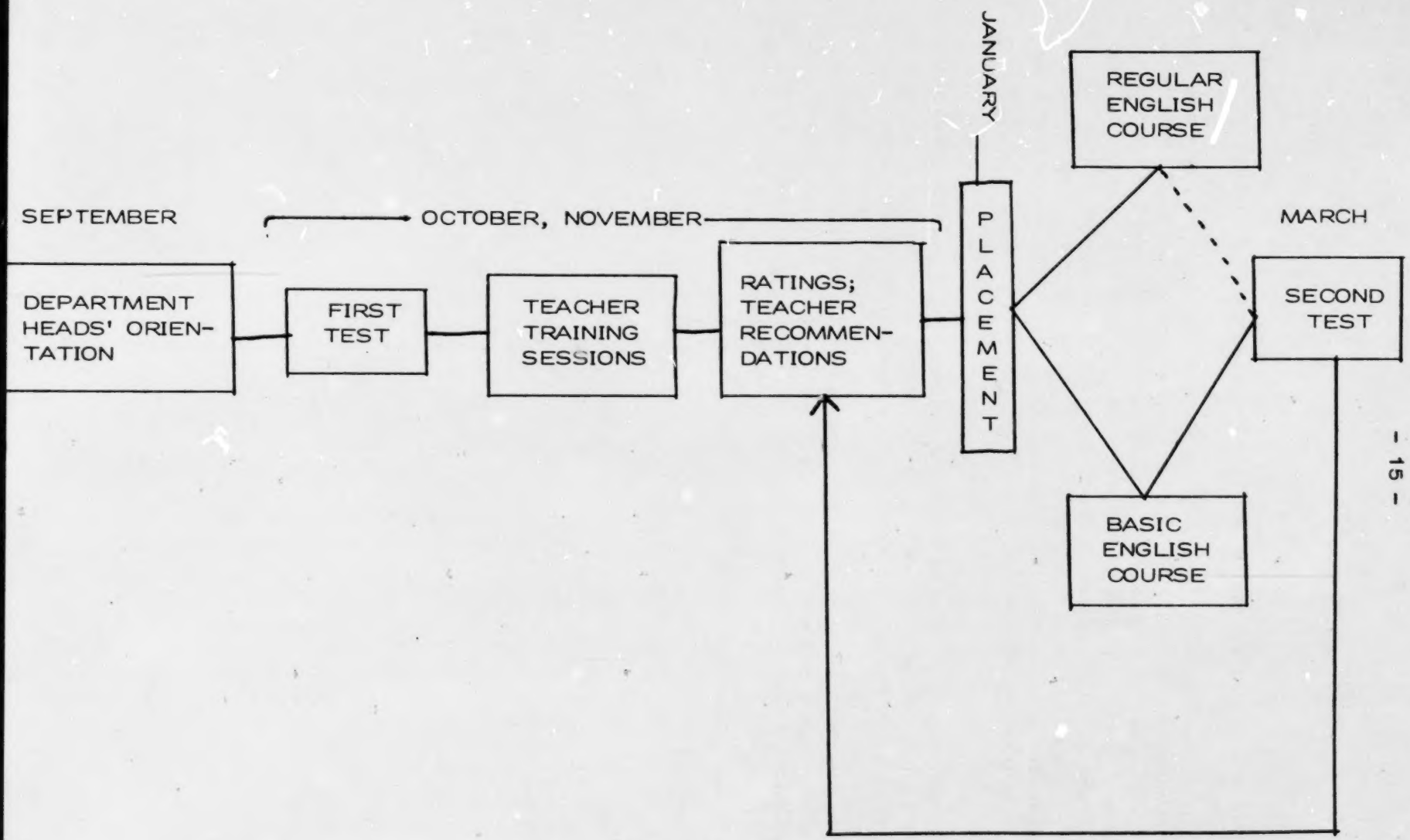
SEPTEMBER
OCTOBER, NOVEMBER
JANUARY
MARCH
DEPARTMENT HEADS' ORIEN-TATION
FIRST TEST
TEACHER TRAINING SESSIONS
RATINGS; TEACHER RECOMMEN-DATIONS
PLACEMENT
REGULAR ENGLISH COURSE
BASIC ENGLISH COURSE
SECOND TEST

Figure 6

scheduling of low-scoring students into mandated remedial classes.

During the month of October, the Project Director conducts one-day Teacher Training Sessions with English teachers at each school. After explanations of program goals and structure, roles of department heads and teachers, and the nature of the entire testing/placement sequence, the teachers rate actual student papers for practice in using the Reader's Scale. Discussions of widely variant ratings (beyond 10 points) on the same paper are held so that teachers can arrive at a concensus about the quality of writing within their school. (An opaque projector is helpful here.)

Teachers are instructed to staple the Reader's Scale to back of the student's paper, below the student's name, grade level, etc. The second reader must avoid the natural temptation to look at the first reader's score before examining the paper. However, the second reader does add his/her rating sum to that of the first reader, entering the Final Score and stapling his/her scale on top of the first Reader's Scale. (See revised Reader's Scale, Figure 7). During the rating process, which

(Figure 7)

Revised

Reader's Scale

(circle one)	Low		Middle		High	
Ideas	2	4	6	8	10	
Organization	2	4	6	8	10	
Wording	1	2	3	4	5	
Flavor	1	2	3	4	5	
						_____
Usage	1	2	3	4	5	
Punctuation	1	2	3	4	5	
Spelling	1	2	3	4	5	
Handwriting	1	2	3	4	5	
						_____
					sum	_____

Second Reader: Place Final Score
(sum of both readers' scores) here: _____

should be completed in three weeks, the department head is responsible for equitable and efficient distribution of the student papers and Reader's Scales to the English teachers.

The Teacher Training Sessions constitute the main expense of the program. English teachers must get a full day of released time, and substitutes or class coverage must be provided. If the local school district does not have one or more central office specialists in English, provision must be made for a highly skilled person to conduct Teacher Training Sessions, schedule the program activities, and monitor it frequently.

Within three weeks after the training session, then, each student's paper should have been rated independently and anonymously by two teachers. The department head and a committee consisting of the principal and one English teacher at each grade level determine appropriate cut-off scores, beneath which students will be required to take specially designed English courses with remedial composition components. Students who pass the test will not be re-tested, but they are instructed in composition on a regular basis in their English classes. If a teacher feels that a student's progress in the regular program is not satisfactory, that teacher can recommend a re-testing in the next available testing cycle.

Even the low-scoring students are not automatically placed in the remedial program. A student's low score is considered in conjunction with the recommendation of the student's regular English teacher. The regular teacher has a developmental view of students' composition skills. If a skilled student happens to be writing poorly on the day of the test, the teacher has on hand other writings that demonstrate the student's usual level of writing skill. When giving the department head a list of students scoring below the cut-off point, the teacher can present recommendations for exclusion of low scorers who ordinarily perform well. Brief explanations of reasons for the exceptional cases should accompany the recommendations. (Recommendations for re-testing students who have already passed the test but have shown inadequate progress in the regular program can be submitted to the department head not later than one month before the next training session.)

Selection of a cut-off point for remedial placement is based on the committee's perception of the overall quality of writing in the school. Since the program is building-centered, there is no necessary correspondence between cut-off points from one school to another. If the level of writing in the school is considered to be extremely low, a higher cut-off point is warranted so that a larger number of students can get remedial help. A school in which most of the students are fairly capable writers will opt for a low cut-off point. High scorers might be channelled into writing-related school activities (school newspapers, school literary journal, peer

editing tasks) or special courses in advanced writing.

A student's regular classroom teacher may or may not choose to award a grade for the student's test paper. But all teachers are responsible for interpreting the papers to their students, and all test papers are kept during the current semester as part of the student's current writing samples. A frequent observation in the first year of the New Orleans program -- consistent with research in feedback and correction -- was that teachers and students felt more satisfaction in schools where feedback was given and maximum classroom use was made of the writing test papers.

In the second semester (January), students recommended for remediation are placed in a course that emphasizes writing -- though not to the total exclusion of literature and oral language skills. The teachers of this course at each grade level should not be grammar-mongers but individuals skilled in use of modern methods and materials in teaching of composition. Late in March, students in the remedial course are tested, and again double ratings are used. A cut-off score is again determined, and individual scores are again considered in conjuction with teacher recommendations.

At the end of the year, the English department and the school principal review the results of the year's testing. The department head reports on the number of students at each grade level passing the test, the number failing, and the number succeeding after being placed in the remedial program. Problems in testing are discussed, and suggestions for improvement made.

In the modified design, the total student population in the school is tested only once. Thereafter, only students entering the school for the first time are tested, along with those who are re-tested in the remedial program. The program's complexity is therefore greatly reduced after the initial schoolwide tests.

The modified design simplified the Diederich system in other ways. Paper shuffling is almost eliminated by abandonment of the Student Identification Slips in favor of student identification on the back of test papers. The longitudinal Student Writing Records and Quartile Charts are not used, because most students are not re-tested in the modified design. Instead, the student's test score is immediately fed back into the instructional program and used directly, along with the classroom teacher's recommendations, for placement purposes.

The modified design is not presented as a desirable program for purposes of individual student placement. Rather, it is an attempt to answer the question, "What do you do with a mandate for testing in composition?" There are other ways of answering this question. For instance, we can and should do things like educating legislators and community groups to the outrages of existing testing systems. In the meantime, though, we need to make the best of a bad situation by devising the most nearly acceptable testing programs that we can. The design presented above is defensible as a good compromise, one that plays on the most positive aspects of evaluation of writing while minimizing the negative effects.

SELECTED BIBLIOGRAPHY ON THE TEACHING AND TESTING OF WRITING

Adler, Richard. "Back to the Basics: Composition." SLATE Starter Sheet, I, 4 (August, 1976). Urbana, Illinois: National Council of Teachers of English, 1976.

Berger, Alan and Blanche Smith, ed. Classroom Practices in Teaching English, 1972-73: Measure for Measure. Urbana, Illinois: National Council of Teachers of English, 1972.

Black, Hillel. The Truth about College Entrance Examinations and Other Standardized Tests. New York: Hart Publishing Co., 1963.

Bordie, John. "Language Tests and Linguistically Different Learners: The Sad State of the Art." Elementary English, 1970 (47) 814-28.

Britton, James, et al. The Development of Writing. London: Macmillan Company, 1975.

Buros, Oscar K. The Seventh Mental Measurement Yearbook. Highland Park, New Jersey: Gryphon Press, 1972.

Cooper, Charles. "Measuring Growth in Writing," in "Research Roundup" column, C. Cooper, ed. English Journal (March, 1975) pp. 111-117.

Daly, John A. and Michael Miller. "The Empirical Development of an Instrument to Measure Writing Apprehension." Research in the Teaching of English, Winter, 1975, IX, pp. 242-249.

Diederich, Paul. Measuring Growth in English. Urbana, Illinois: National Council of Teachers of English, 1974.

Fagan, William, et al. Measures for Research and Evaluation in the English Language Arts. Urbana, Illinois: NCTE/ERIC, 1975.

Godschalk, Fred, et al. The Measurement of Writing Ability. New York: College Entrance Examination Board, 1966.

Grommon, Alfred, ed. Reviews of Selected Published Tests in English. Urbana, Illinois: National Council of Teachers of English, 1976.

Hoffman, Banesh. The Tyranny of Testing. n.p.: Crowell-Collier Press, 1962.

Houts, Paul H., ed. National Elementary Principal, LIV, 6 (Testing Issue: "The Scoring of Children: Standardized Testing in America" July/August, 1975).

Loban, Walter. Language Development. Urbana, Illinois: National Council of Teachers of English, 1976.

Mellon, John C., et al. National Assessment and the Teaching of English. Urbana, Illinois: National Council of Teachers of English, 1975.

Monroe, Betty. "Review: The Measurement of English." in "Professional Publications" column, C. Suhor, ed. English Journal (March, 1975), pp. 99-100.

National Assessment of Educational Progress. "Report No. 3: 1969-1970 Writing: National Results." Denver, Colorado: Education Commission of the States, 1970.

National Assessment of Educational Progress. "Report No. 5: 1969-1970 Writing." Denver, Colorado: Education Commission of the States, 1971.

National Assessment of Educational Progress. "Writing Mechanics, 1969-1974." Report No. 05-W-01. Denver, Colorado: Education Commission of the States, 1975.

National School Board Association. Standardized Achievement Testing. Washington, D.C.: NSBA, 1977.

Northcutt, Norvell. Adult Functional Competency: A Summary." Austin, Texas: University of Texas at Austin, 1975.

O'Hare, Frank. Sentence Combining: Improving Student Writing Without Formal Grammar Instruction. Urbana, Illinois: National Council of Teachers of English, 1973.

Purves, Alan. "Minimal Competencies and Measures of Competence." SLATE Starter Sheet, II, 7 (December, 1976). Urbana, Illinois: National Council of Teachers of English, 1976.

Purves, Alan, et al. Common Sense and Testing in English. Urbana, Illinois: National Council of Teachers of English, 1975.

Ruth, Leo. "Standardized Testing: How to Read the Results." SLATE Starter Sheet, II, 5 (March, 1977). Urbana, Illinois: National Council of Teachers of English, 1977.